멀티모달 대화형 감정인식을 위한 마트료시카 표현 학습과 모달리티별 지식 증류 기법

나영진, 이성규, 신종원
광주과학기술원
{nayoungjin920, lsqjin2022}@gm.gist.ac..kr, jwshin@gist.ac.kr

Matryoshka Representation Learning with Modality-wise Knowledge Distillation for Multimodal Emotion Recognition in Conversation

Young-jin Na, Seonggyu Lee and Jong Won Shin
Gwangju Institute of Science and Technology, Korea

요약

본 논문은 대화 속 각 발화에 대한 감정 레이블을 분류하는 멀티모달 대화형 감정인식 모델에 Matryoshka Representation Learning (MRL)과 Modality-wise Knowledge Distillation (MKD)을 적용하여 감정 예측 부분에서의 성능 향상을 달성하고자 한다. 대화형 감정인식에서는 발화의 순간적인 억양 변화와 대화 전반의 맥락을 동시에 포착해야 하나, 기존 단일 크기 임베딩 기반 모델은 이러한 다중 수준의 의미 정보를 모두 보존하기 어렵다. 이를 해결하기 위해 기존 임베딩 공간을 다중 해상도로 학습하는 MRL을 적용하여 짧은 단서와 대화의 맥락 정보를 효과적으로 반영하였다. 또한 멀티모달 구조에서는 모달리티별로 노이즈, 정보량, 중요성이 달라 특정 모달리티가 과도하게 최적화될 경우, 최적의 성능을 제공하지 못할 수 있다. 본 연구에서는 단일모달 감정인식 모델을 활용한 MKD를 적용하여 모달리티 간 불균형을 완화하고 감정인식 성능을 향상시켰다.

Ⅰ. 서론

감정은 인간의 본질을 이루는 중요한 요소로, 이를 표현하는 주요 방법은 대화이다. 따라서 대화 속 감정을 정확하게 이해하는 것은 중요하며 이는 인공지능형 대화, 심리치료 등 다양한 응용 분야에서 핵심 과제로 여겨진다. 멀티모달 대화형 감정인식은 대화 속의 발화(텍스트, 오디오)와 시각적(비디오) 정보를 이용하여 화자의 감정을 식별하는 것을 목표로 한다. 최근 높은 성능을 보인 모델 [1]은 멀티모달 융합 이전의 모달 간 정렬 과정과 모달 내 노이즈 정보를 고려하는 재정렬 마스킹 그래프 학습 기법을 제안하였다.

그래프 뉴럴 네트워크 모델은 대화 상황을 노드와 엣지로 구성된 그래프 구조로 변환하여 발화 간의 상호작용 관계를 효과적으로 학습한다. 이는 복잡한 관계망 속에서 정보가 전달되는 소셜 네트워크 분석이나 분자 구조 예측과 같은 분야에서 처음 소개되어 우수한 성능을 보여왔다. 단순히 발화의 개별 특징만을 고려하는 방식과 달리, 대화 상황에서는 발화 간의 상호 연결성이 감정적 의미에 중요한 영향을 미치므로 그래프 뉴럴 네트워크 모델은 인접 발화 간의 맥락적 의존성을 보존하면서 감정인식을 수행할 수 있다. 그러나 그래프 뉴럴 네트워크를 통해 생성된 기존 단일 크기 임베딩 기반 모델은 여전히 대화 속 발화의 다중 수준 의미 정보를 모두 보존하기 어렵다는 한계가 있다. 또한 멀티모달 구조에서는 모달리티별 정보량, 노이즈, 중요성이 달라 특정 모달리티가 과도하게 최적화될 경우 전체 성능이 저하될 수 있다.

이러한 문제를 해결하기 위해 본 연구는 (1) 임베딩을 다중 해상도로 학습하는 MRL [2]을 통해 다양한 수준의 의미 정보를 포착하고, (2) 단일모달 모델을 활용하는 MKD [3]를 통해 모달리티 간 불균형을 완화하는 방법을 제안한다. 이를 통해 멀티모달 대화형 감정인식 모델의 표현력을 강화하고, 대화형 감정인식 벤치마크 데이터셋인 IEMOCAP [4]과 MELD [5]에서 최신 방법들을 뛰어넘는 성능을 보였음을 입증하고자 한다.

Ⅱ. 본론

본 논문에서는 최신 그래프 뉴럴 네트워크 기반의 대화형 감정인식 모델 (Masked Graph Learning with Recurrent Alignment for Multimodal Emotion Recognition in Conversation, MGLRA) [1]에 대해 제안하는 알고리즘을 실험한다.

1. Matryoshka Representation Learning (MRL)

$d \in \mathbb{N}$에 대해 표현 크기의 집합 $M \subset [d]$를 고려하면, MRL은 입력 도메인 χ의 데이터 포인트 x에 대해 d차원 표현 벡터 $z \in \mathbb{R}^d$를 학습한다. 또한 모든 $m \in M$에 대한 임베딩 벡터의 앞쪽 m차원 $z_{1:m} \in \mathbb{R}_m$는 독립적으로 데이터 포인트 x의 전이 가능하고 범용적인 표현으로 사용할 수 있다. 학습 가능한 파라미터 θ_F로 최적화된 MGLRA를 $F(\cdot;\theta_F): \chi \to \mathbb{R}^d$로 정의하면 표현 벡터 z는 $z := F(x;\theta_F)$로 정의된다. 입력데이터 $x_i \in X$와 x_i에 대한 라벨 $y_i \in [L]$가 주어졌을 때, MRL은 각 중첩된 차원 $m \in M$에 대해 empirical risk minimization (ERM)을 사용하여 각 차원별로 고유한 선형 분류기 $W^{(m)} \in \mathbb{R}^{L \times m}$을 학습한다. 즉, 다음과 같은 최적화 문제를 푼다:

$$\min_{W^{(m)}_{m\in M},\, \theta_F} \frac{1}{N} \sum_{i \in [N]} \sum_{m \in M} \mathbb{L}\left(W^{(m)} \cdot F(x_i;\theta_F)_{1:m}; y_i\right).$$

여기서 $\mathbb{L}$은 multi-class softmax cross-entropy이며, 각 표현 크기마다 독립적으로 분류기를 학습하면서도 $O(\log_2(d))$개의 축소된 표현을 학습하면서 대화 속 발화의 다양한 해상도의 표현을 학습한다. 또한 공통 가중치 $W \in \mathbb{R}^{(L \times d)}$에 대해 $W^{(m)} = W_{1:m}$로 정의하여 모든 선형 분류기를 공유하는 MRL-E를 통해 메모리 효율성을 향상시킬 수 있다.

2. Modality-wise Knowledge Distillation (MKD)

MGLRA내의 모든 단일모달 인코더 E_k의 최적화를 돕기위해 각 모달리티 $k \in K$(텍스트, 오디오, 비디오)에 대한 단일모달 모델 $\overline{U_k}$을 학습하

여 pseudo label을 생성하고 단일모달 모델 U_k을 지도한다. 지식 증류 과정에서 선생 모델인 $\overline{U_k}$와 학생 모델인 U_k은 다음과 같이 표현된다:

$$\overline{U_k} = \overline{C_k}(\overline{E_k}(x_k)),\ U_k = C_k(E_k(x_k)).$$

또한 각 모델로부터 얻은 레이블 벡터 $\overline{y_k}$, y_k는 다음과 같이 정의된다:

$$\overline{y_k} = \text{softmax}(\overline{U_k}),\ y_k = \text{softmax}(U_k).$$

따라서 MKD의 손실함수는 다음과 같이 정의된다:

$$\mathbb{L}(\overline{y_k}, y_k)_{k=1}^{K} = \sum_{k=1}^{K} \mathbb{L}_{CE}(\overline{y_k}, y_k)$$

각 단일모달 인코더는 선생 모델로부터 지도받으며 점진적으로 해당 지식에 맞춰 업데이트되어 모달리티 간 최적화 불균형을 완화한다.

3. MRL(-e) with MKD

제안하는 시스템 구조는 그림 1과 같다. 전반적인 모델 구조는 MGLRA와 동일하며 그래프 뉴럴 네트워크를 통해 생성된 각 발화에 대한 표현은 MRL을 통해 여러 크기의 벡터로 나눠지고 각각의 선형 분류기를 학습하게 된다. 또한 MKD를 통해 MGLRA내의 각 단일모달 인코더를 최적화하여 모달리티 간의 불균형 문제를 해결한다.

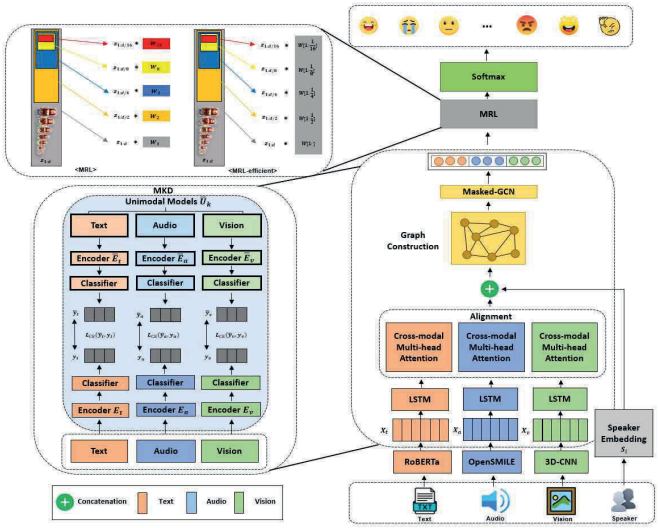

그림 1. 제안하는 시스템 구조

4. 실험 세팅 및 결과

대화형 감정인식 벤치마크 데이터셋인 IEMOCAP과 MELD를 사용하여 모델의 학습 및 평가를 진행하였다. IEMOCAP은 감정이 두드러지는 대화 상황을 가정해 남자와 여자가 연기를 하는 데이터셋이며 153개의 대화, 10명의 화자, 7433개의 발화로 구성되어있으며 감정 레이블은 행복, 슬픔, 중립, 분노, 흥분, 좌절까지 총 6개로 분류한다. MELD는 Friends TV 프로그램에서 수집되었으며 1433개의 대화, 13708개의 발화로 구성되어있다. 감정 레이블은 중립, 놀람, 두려움, 슬픔, 즐거움, 역겨움, 분노까지 총 7개로 분류된다. 성능 측정을 위해 정확도(Acc)와 F1-score를 사용하였으며 결과는 표 1과 같다.

표 1에서 MGLRA에 MRL, MRL-e, MKD를 적용했을 때 전반적인 성능 향상을 보였다. MRL with MKD는 MGLRA보다 IEMOCAP에서 3.2%의 정확도, 3.05%의 F1-score가 향상되었으며 MELD에서 0.84%의 정확도, 0.91%의 F1-score가 향상되었다. 또한 MRL-e with MKD는 MGLRA보다 IEMOCAP에서 3.63%의 정확도, 3.34%의 F1-score가 향상되었으며 MELD에서 0.84%의 정확도, 0.96%의 F1-score가 향상되어 가장 높은 정확도와 F1-score를 달성하였다. 제안하는 방법은 멀티모달 대화형 감정인식에서 임베딩의 다중 해상도 학습과 모달리티 간 불균형 완화를 동시에 달성함으로써 기존 모델의 한계를 극복하고 최신 기법 대비 우수한 성능을 입증하였다.

Method	IEMOCAP		MELD	
	Acc(%)	F1-score	Acc(%)	F1-score
*MGLRA	68.52	68.26	66.86	65.33
MRL	70.49	70.19	67.20	66.25
MRL-e	70.18	69.92	67.24	66.21
MKD	71.53	70.77	66.82	66.07
MRL with MKD	71.72	71.31	**67.70**	66.24
MRL-e with MKD	**72.15**	**71.60**	**67.70**	**66.29**

표 1. IEMOCAP과 MELD에서의 실험 성능 결과 (*모델 재구현 결과)

Ⅲ. 결론

본 논문에서는 멀티모달 대화형 감정인식의 성능을 향상시키기 위해 MRL과 MKD를 결합하는 새로운 방법을 제안하였다. MRL은 임베딩을 다중 해상도로 학습함으로써 발화의 순간적 단서와 대화의 맥락을 동시에 반영할 수 있도록 하였으며, MKD는 단일모달 별 선생 모델의 지식을 활용하여 모달리티 간 최적화 불균형을 완화하였다. IEMOCAP과 MELD 두 가지 벤치마크 데이터셋에서 제안한 방법은 최신 그래프 뉴럴 네트워크 기반 모델(MGLRA) 대비 정확도와 F1-score 모두에서 의미 있는 성능 향상을 달성하였다. 특히 MRL-e with MKD는 두 데이터셋 모두에서 가장 높은 성능을 기록하며, 다중 해상도 표현 학습과 모달리티 간 지식 증류의 효과성을 입증하였다. 이러한 결과는 멀티모달 대화형 감정인식에서 (1) 다중 수준 의미 정보의 보존과 (2) 모달리티 간 불균형 문제 해결이라는 두 가지 핵심 과제를 동시에 해결할 수 있는 방안을 제시하였다.

ACKNOWLEDGMENT

이 논문은 2025년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (RS-2022-II220989, 다화자 동시처리를 위한 인공지능 기반 대화 모델링 기술 개발)

참 고 문 헌

[1] Meng, Tao, et al. “Masked graph learning with recurrent alignment for multimodal emotion recognition in conversation.” *IEEE/ACM Transactions on Audio, Speech, and Language Processing* (2024).

[2] Kusupati, Aditya, et al. “Matryoshka representation learning.” *Advances in Neural Information Processing Systems 35* (2022): 30233-30249.

[3] Lee, S., Ahn, Y., Shin, J. W. “Multimodal Emotion Recognition Using Modality-wise Knowledge Distillation.” *Sensors,* 2025 (submitted)

[4] Busso, C., Bulut, M., Lee, et al. “IEMOCAP: Interactive emotional dyadic motion capture database.” *Language resources and evaluation,* vol. 42, no. 4, pp. 335-359, 2008.

[5] Poria, S., Hazarika, D., et al. “Meld: A multimodal multi-party dataset for emotion recognition in conversations.” in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics,* 2019, pp.527-536